Research Article Impact Factor: 0.621 ISSN: 2319-507X
Arvindkumar Mishra, IJPRET, 2014; Volume 2 (9): 434-440 IJPRET


INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

# A SURVEY ON EXTRACTING INFORMATION TO KNOWLEDGE DISCOVERY IN DATABASE

**PROF. ARVINDKUMAR D. MISHRA**

Dept. Of Computer Science, Vidya Bharati Mahavidyalya Amravati

**Accepted Date: 27/02/2014 ; Published Date: 01/05/2014**

**Abstract:** Data mining and Knowledge discovery has several important application areas. Data Mining and knowledge discovery have been topics considered at many AI, database and Statistical conferences. Knowledge discovery generally refers to the process of identifying valid, novel and understandable patterns. Knowledge discovery from large databases, often called data mining, refers to the application of the discovery process on large databases or datasets. In this paper, we provide an overview of knowledge discovery tasks and process to extract data from database. Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world Applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery and current and future research directions in the field.

**Keywords:** Knowledge discovery in databases, data mining, surveys, Data, Information.

.

**Corresponding Author: MR. NAVED RAZA Q. ALI**

**Access Online On:**

www.ijpret.com

**How to Cite This Article:**

Arvindkumar Mishra, IJPRET, 2014; Volume 2 (9): 434-440

***PAPER-QR CODE***

## INTRODUCTION

*Knowledge Discovery in Databases* (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. *Data Mining* (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding Phenomena from the data, analysis and prediction. Data in processing state is called information. The final output cloth is the knowledge. A person having more knowledge is called highly intelligence person. Data store in a data base where as knowledge store in a knowledge base. Across a wide diversity of fields, data are being collected and accumulated at a remarkable speed. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases[1][2][3].

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [3]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [1].

Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line [5]. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?" This paper provides an introduction to the basic technologies of data mining [2]. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

## 2. Why Do We Need KDD

Knowledge Discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data With the emphasis on collecting data increasing

around the world, there is an urgent need for a new generation of different techniques, methods and algorithms to assist researchers, analysts, decision makers and managers in extracting useful patterns from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD). KDD has evolved from interaction and cooperation among such different fields as machine learning, pattern recognition, database, statistics, artificial Intelligence, knowledge representation, and knowledge acquisition for intelligent systems[1][5].

**3. Data Mining Approach**

Data mining approaches can be used for prediction and description purposes, there is no sharp distinction between prediction and description, but many of the researchers in this field believe that the KDD approach focuses more on description than prediction. However, the knowledge extracted from a given level of KDD could be applied in another project. The goal of prediction and description is achieved by applying different data mining methods [3]. We mention a few popular methods:

- Classification: This basically is the process of finding a function which maps the data to recognized classes.
- Regression: process of finding a function to map the data to a real valued prediction variable and discovery the relationships among variables [6].
- Clustering: unsupervised identifying a set of cluster in a data set.
- Summarization: finding an abstract level of knowledge description.
- Association Analysis: is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness
- Anomaly Detection :(or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset
- Sequence analysis: refers to the process of subjecting or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. Methodologies used include sequence alignment, searches against biological databases, and others
- Time series: analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data

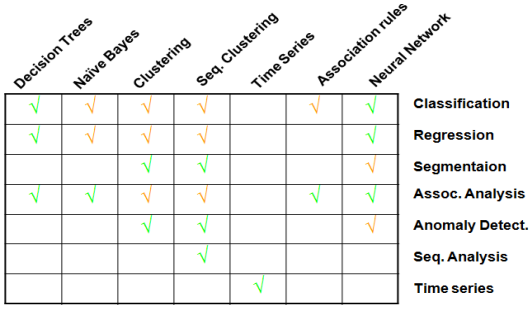


**Figure 1 Data mining approaches analysis**

## 4. The Knowledge Discovery Process

There is still some confusion about the terms Knowledge Discovery in Databases and data mining. Often these two terms are used interchangeably. We use the term KDD to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [4].

We often see data as a string of bits, or numbers and symbols, or "objects" which we collect daily. Information is data stripped of redundancy, and reduced to the minimum necessary to characterize the data. Knowledge is integrated information, including facts and their relations, which have been perceived, discovered, or learned. Knowledge can be considered data at a high level of abstraction and generalization

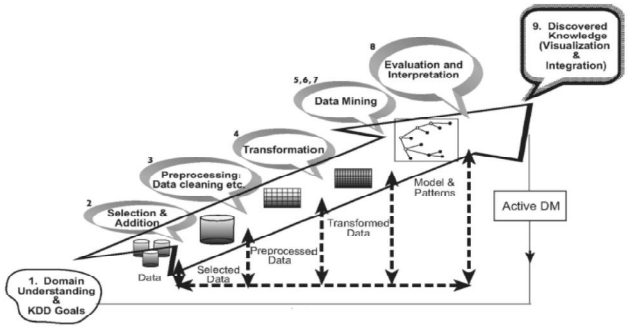


**Figure 1.1 The Process of Knowledge Discovery in Databases.**

The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. Then the loop is closed – the Active Data Mining part As a result, changes would have to be made in the application this closes the loop, and the effects are then measured on the new data repositories, and the KDD process is launched again[1][2].

**4.1 Developing an understanding of the application domain**

This is the initial preparatory step. It prepares the scene for understanding what should be done with the many decisions (about transformation, algorithms, representation, etc.).

**4.2 Selecting and creating a data set on which discovery will be performed.**

Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From this respect, the more attributes are considered, the better. On the other hand, to collect, organize and operate complex data repositories is expensive and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. This starts with the best available data set and later expands and observes the effect in terms of knowledge discovery and modeling[1][3].

**4.3 Preprocessing and cleansing.**

In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. It may involve complex statistical methods or using a Data Mining algorithm in this context. For example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors.

**4.4 Data transformation.**

In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction, this step can be crucial for the success of the entire KDD project, and it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. In marketing, we may need to consider effects beyond our control as well as. Efforts and temporal issues. However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed. Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed.

**4.5 Choosing the appropriate Data Mining task.**

We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta-learning for the particular set of available data.

**4.6 Choosing the Data Mining algorithm.**

Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers)

**4.7 Employing the Data Mining algorithm.**

Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the Minimum number of instances in a single leaf of a decision tree.

**4.8 Evaluation.**

In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Here we consider the preprocessing steps with respect to their effect on the Data Mining algorithm results and repeating from there). This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage. The last step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining

**4.9 Using the discovered knowledge.**

We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we may make changes to the system and measure the effects. Actually the success of this step determines the effectiveness of the entire KDD process. There are many challenges in this step, such as loosing the "laboratory conditions" under which we have operated. For instance, the knowledge was discovered from a certain static snapshot (usually sample) of the data, but now the data becomes dynamic [2].

## 5. CONCLUSION

In conclusion, I presented some definitions of basic notions in the KDD field. My primary aim was to clarify the relation between knowledge discovery and data mining. I provided an overview of the KDD process and basic data-mining methods. Given the broad spectrum of data-mining methods and algorithms, our overview is inevitably limited in scope: There are many data-mining techniques, particularly specialized methods for particular types of data and domain. Although various algorithms and applications Might appear quite different on the surface, it is not uncommon to find that they share many common components. Understanding data mining and model induction at this component level clarifies the task of any data- mining algorithm and makes it easier for the user to understand its overall contribution and applicability to the KDD process.

## 6. REFERENCES

1. Knowledge Discovery in Databases: An Overview William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus AI Magazine Volume 13 Number 3 (1992) (© AAAI)

2. From Data Mining to Knowledge Discovery in Databases Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth AI Magazine Volume 17 Number 3 (1996)

3. A SURVEY OF DATA MINING AND KNOWLEDGE DISCOVERY SOFTWARE TOOLS-Michael Goebel Department of Computer Science University of Auckland Private Bag 92019, Auckland New Zealand mgoebel@cs.auckland.ac.nzSIGKDD Explorations. Copyright 1999 ACM SIGKDD, June 1999.

4. Data Mining Techniques: A Tool For Knowledge Management System In Agriculture -Latika Sharma, Nitu Mehta INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 1, ISSUE 5, JUNE 2012 ISSN 2277-8616

5. Data Warehousing and Knowledge Discovery: A Chronological View of Research Challenges Tho Manh Nguyen1, A Min Tjoa 1, and Juan Trujillo A Min Tjoa and J. Trujillo (Eds.): DaWaK 2005, LNCS 3589, pp. 530 – 535, 2005. © Springer-Verlag Berlin Heidelberg 2005

6. Mining knowledge using Decision Tree Algorithm: Mrs. Swati .V. Kulkarni International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011 1 ISSN 2229-5518

7. Web references www.wikipedia.com, www.google.com